Psychological Topics, 28 (2019), 1, 1-20

Original Scientific Paper UDC: 159.955

doi:https://doi.org/10.31820/pt.28.1.1

Heuristic Cues for Meta-Reasoning Judgments: Review and Methodology

Rakefet Ackerman

Technion – Israel Institute of Technology, Haifa, Israel

Abstract

Metacognitive research aims to explain how people regulate their effort when performing cognitive tasks, to expose conditions that support reliable monitoring of chance for success, and to provide a basis for developing improvement guidelines. The essence of the domain is that monitoring drives control: people continually self-assess their chance for success before, during, and after performing a cognitive task, and use these judgments to guide their effort-allocation decisions (e.g., whether to reconsider an answer option, change strategy, seek help, or give up). Thus, factors that underlie metacognitive judgments affect the efficiency with which people perform cognitive tasks. This paper focuses on meta-reasoning – the monitoring and control processes that apply to reasoning, problem-solving, and decision-making tasks. So far, relatively little is known about heuristic cues used for inferring meta-reasoning judgments. This paper reviews the known heuristic cues and offers methodological guidelines for a critical reading of existing research and for designing high-quality studies that will advance this important domain.

Keywords: metacognition, reasoning, problem solving, metacognitive monitoring, heuristic cues

Introduction

Performing cognitive tasks, such as problem solving or text learning, requires activation of various cognitive operations for each task item (e.g., solving each in a series of puzzles, or studying a paragraph within a lengthy text). These operations include retrieval of relevant prior knowledge, representation of the various task components, and whatever specific steps (e.g., induction, deduction, abstraction, application of quantitative rules, memory retrieval, etc.) are required for the given task (Butler & Winne, 1995). Performing such tasks also involves a parallel set of

✉ Rakefet Ackerman, Faculty of Industrial Engineering & Management, Technion – Israel Institute of Technology, Haifa, 3200003, Israel. E-mail: *Ackerman@ie.technion.ac.il*

This work was supported by the Israel Science Foundation [grant No. 234/18].

metacognitive processes, including setting a goal for each item, monitoring progress toward that goal, and directing one's effort in accordance (see Ackerman & Thompson, 2017a; Bjork, Dunlosky, & Kornell, 2013, for reviews). The metacognitive research approach within cognitive psychology, focuses on exposing bases for people's monitoring and the actions taken in response in terms of effort allocation and time management (e.g., deciding to reconsider an answer, change strategy, seek help, or cease investing effort).

Monitoring can be expressed by many types of judgments, which people are assumed to make spontaneously before, during, and after performance of any cognitive task. These judgments include an initial judgment of whether the task is doable, followed by assessments of initial outcomes, ongoing progress, and the chance for success of a chosen response. The basic principle is that monitoring drives control over effort allocation (Nelson & Narens, 1990). For example, judgments of learning (JOLs) provided after memorizing words were found to be causally related to restudy choices (Metcalfe & Finn, 2008). In math exercises ($12\times14=?$), feeling of knowing (FOK) about the exercise's components guided attempts to retrieve a known solution (Reder & Ritter, 1992). Feeling of rightness (FOR), a judgment that applies to initial intuitive solutions (the solution that jumps to mind), was found to be associated with reconsideration time and likelihood of changing the solution (Thompson, Prowse Turner, & Pennycook, 2011). Initial judgments of solvability assessing whether Raven Matrices are solvable (vs. mixed figures without underlying rules) provided after a brief glance in the matrices, were found to predict the time people invest later in attempting the problems, above and beyond other potential cues (Lauterman & Ackerman, in press). Similarly, confidence was found to predict information-seeking in decision-making contexts (Desender, Boldt, & Yeung, 2018). Thus, as long as the relevant judgment is reliable, people have a solid basis for making effective control decisions (that is, decisions regarding effort allocation). Unreliable judgments lead to bad decisions (see Bjork et al., 2013, for a review). For instance, people who feel overconfident when performing a challenging task are likely to cease investing effort too early, when in fact they should attempt to improve their chance for success by allocating more time to the task (e.g., Ackerman & Goldsmith, 2011).

Metacognitive judgments are known to be based on heuristic cues (Koriat, 1997; see Dunlosky & Tauber, 2014 for a review). That is, people cannot directly "read" the quality of their own cognitive processing, but instead apply *cue utilization* – they base their metacognitive judgments on information drawn from the task, the environment, or their own subjective experience. Based on these heuristic cues people infer their own chance for success at any given moment. The predictive accuracy of metacognitive judgments depends on *cue diagnosticity* – the diagnostic value of the heuristic cues that underlie them.

Most research dealing with heuristic cues for judgments has been done with memorization and knowledge retrieval tasks, under the meta-memory research

domain (e.g., Koriat, 2012; Metcalfe & Finn, 2008), and with learning from texts, under the meta-comprehension research domain (see Wiley et al., 2016, for a review). In recent years, a growing body of literature has begun to consider the heuristic cues which underlie metacognitive monitoring in the context of reasoning, problem solving, and decision making, under the meta-reasoning framework (Ackerman & Thompson, 2015, 2017a, 2017b). While most principles are common across task domains, some heuristic cues have been found to affect metacognitive judgments differently across domains (e.g., effects of font readability on metacognitive judgments in reasoning, Thompson, Prowse Turner et al., 2013; vs. in memorizing, Undorf, Söllner, & Bröder, 2018; Undorf & Zimdahl, 2019). In this review I focus on the meta-reasoning context.

A wide-angle view of the metacognitive literature suggests three levels of heuristic cues for metacognitive judgments (see Box 1). Classic meta-memory research has focused mostly on the last level – people's momentary subjective experience when encountering each item (e.g., a word pair to be memorized). Bringing to the fore the other two levels highlights that people are quite sophisticated in integrating self-perceptions and task characteristics in their judgments, along with a variety of momentary experiences (e.g., Bajšanski, Žauhar, & Valerjev, in press; Koriat, Ma'ayan, & Nussinson, 2006; Thompson, Pennycook, Trippas, & Evans, 2018; Undorf et al., 2018). This complex inference process seems to develop throughout childhood and matures only towards adulthood (Koriat, Ackerman, Adiv, Lockl, & Schneider, 2014; van Loon, Destan, Spiess, de Bruin, & Roebers, 2017).

Level 1: Self-Perceptions

Self-perceptions refer to a person's beliefs about his/her own traits, abilities, or knowledge, either in general or with respect to a given task type or domain. For example, test anxiety and math anxiety derive from self-doubt about a particular task type (test-taking) or domain (mathematics), respectively, regardless of the particular task one might face at a given moment (e.g., Morsanyi, Busdraghi, & Primi, 2014). Another example is need for cognition, which reflects the extent to which a person enjoys (or dislikes) effortful cognitive activities (Cacioppo & Petty, 1982). Data regarding self-perceptions are typically collected through self-report questionnaires. One important aspect of self-perceptions is confidence in one's ability to succeed at a given task. In metacognitive research, meta-reasoning included, the main approach to assessing confidence is through item-level confidence ratings (i.e., one or more ratings collected for each item in a task), rather than through self-reported confidence about a global task type or domain, as the detailed confidence guides effort allocation for each item, as reviewed above. The means of item-level confidence across items (e.g., exam questions) can be calculated to produce an overall appraisal of a person's item-level confidence when performing a task. Overall confidence assessed in this way has been found to be associated with various self-perceptions (math anxiety, Legg & Locker, 2009; analytic-thinking disposition, Pennycook, Ross, Koehler, &

Fugelsang, 2017; self-reported thinking style, Prowse Turner & Thompson, 2009; English, math, academic, and memory self efficiency and self concepts, Stankov, Lee, Luo, & Hogan, 2012) and is accounted a stable trait (Jackson & Kleitman, 2014; Stankov, Kleitman, & Jackson, 2014). Notably, though, recent findings with perceptual, knowledge, reasoning, and emotion identification tasks suggest that the discrimination between correct and wrong responses is more malleable than the global confidence and overconfidence levels, especially when considering experimental designs in which each individual performs particularly diverse tasks (Ais, Zylberberg, Barttfeld, & Sigman, 2016; Dentakos, Saoud, Ackerman, & Toplak, in press).

Box 1. Levels of Heuristic Cues for Metacognitive Judgments

A review of the metacognitive literature reveals three levels of cues for metacognitive judgments and interactions among them:

Level 1: Self-perceptions

Overall assessment of one's own qualities in a given task domain.

Examples:

- "I am good/bad at this type of task"
- "I have a good/bad memory for details"
- Domain knowledge (e.g., level of expertise)
- Relevant acknowledged personality traits (e.g., test anxiety, need for cognition).

Level 2: Task characteristics

Information and beliefs about factors affecting performance in a task as a whole.

Examples:

- Test type (e.g., open-ended vs. multiple-choice test format, memory for details vs. high-order comprehension)
- Time frame (pressured vs. loose)
- Environment (e.g., computer vs. paper, indoors vs. outdoors, home vs. classroom)
- With/without training or feedback
- Instructions (e.g., emphasizing speed vs. accuracy)

Level 3: Momentary experiences

Item-level indications of chance for success based on momentary subjective experience before, during, and after attempting any task item (e.g., answering a question in an exam).

Examples:

Fluency (perceived ease of processing), consensuality, accessibility, cardinality, familiarity, concreteness, coherence, pronounceability

A consistent finding across many studies and methodologies is that lower achievers tend to be less confident than higher achievers. However, lower achievers typically do not acknowledge just how low is their actual success rate, meaning that their confidence should in fact be even lower. Consequently, lower achievers are more overconfident than higher achievers (see Figure 1 for an example). The greater overconfidence of lower achievers compared to higher achievers is a manifestation of the classic Dunning-Kruger effect (see Pennycook et al., 2017). Identifying the heuristic cues that lead lower achievers astray is an intriguing and important focus of metacognitive research that so far is understudied (e.g., Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008).

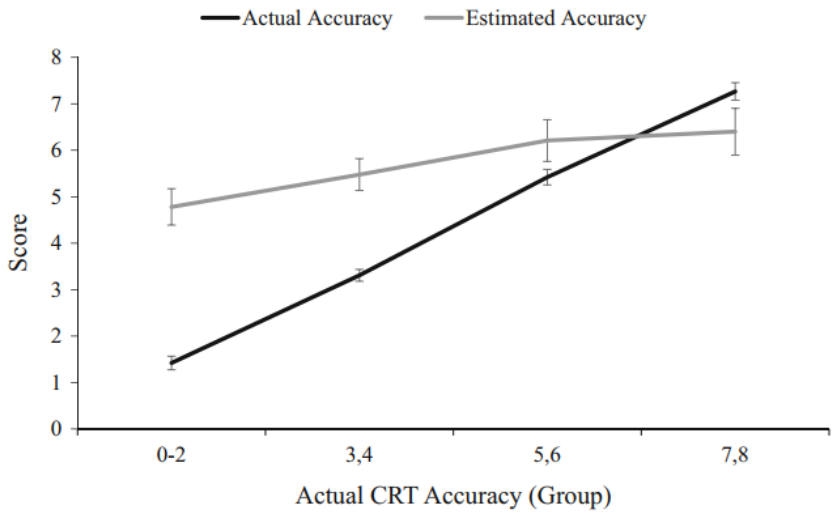

Figure 1. Example of the Dunning-Kruger effect, adapted from Pennycook et al. (2017). The graph shows participants' estimated and actual accuracy in an 8-item reasoning test (a variation of the Cognitive Reflection Test, or CRT) as a function of correct answers per participant.

From a methodological perspective, mean confidence judgments based on item-level ratings have a number of uses in metacognitive research. When item-level confidence ratings are collected on Likert scales (e.g., 1 = *sure to be wrong*, 7 = *absolutely confident*), they can be used to compare mean confidence across conditions or groups (e.g., Thompson et al., 2018). Likert scales also allow examining correlations between various responses of the same individual and across individuals (e.g., Stupple, Ball, & Ellis, 2013). More precise confidence ratings, such as those collected as a percentage (0-100% confidence) or ratio (e.g., number of items judged to be correct relative to the total number of items), allow comparisons between mean confidence and actual success rates (e.g., Pennycook et al., 2017; Sidi,

Shpigelman, Zalmanov, & Ackerman, 2017). Such absolute ratings by percentage or ratio allow examining calibration in terms of over- and underconfidence.

Level 2: Task Characteristics

The second set of cues is information about characteristics of the tasks (see examples in Box 1). These characteristics may affect performance or may be unduly thought to affect it. Notably, people tend to under- or overestimate these characteristics' effect on performance based on naïve theories (Mueller & Dunlosky, 2017). For instance, when people are allowed first-order experience of solving a task before being asked to assess their chance for success, they tend to pay more attention to item-level cues and underestimate the effect of task characteristics on their performance in both memory (recognition vs. recall test format; Touron, Hertzog, & Speagle, 2009) and reasoning contexts (open ended vs. multiple choice test format, Ackerman & Zalmanov, 2012, see more details and Figure 2 below; solving syllogisms with vs. without training, Prowse Turner & Thompson, 2009). We can see that it is the presence of item-level cues which interferes with cues based on task characteristics, because the latter have a stronger effect when participants provide judgments without an opportunity to attempt the tasks themselves (e.g., when assessing the difficulty of finding an answer presented by others) than when they can use their first-order experience (Kelley & Jacoby, 1996; Mitchum & Kelley, 2010).
In one line of research, Ackerman and colleagues (Ackerman & Lauterman, 2012; Sidi et al., 2017) examined two task characteristics in both text learning and problem-solving tasks: time frame (working under time pressure versus a loose time frame) and medium (encountering the task on a computer screen versus on paper). They found that on paper participants performed equally well in both time frames (with and without time pressure), while participants working on screens performed as well only under free time regulation. Success rates for participants working on screen under time pressure were significantly lower than in all other conditions (see Delgado, Vargas, Ackerman, & Salmerón, 2018, for a meta-analysis that shows this pattern to be robust). Notably, however, the interactive effect of time frame and medium on performance was not reflected in participants' metacognitive judgments. Overall, the judgments were a little lower under time pressure relative to free time, regardless of the actual performance difference between the time frames; and the metacognitive judgments did not capture the performance difference between the media under time pressure. Similar findings were reported by Shynkaruk and Thompson (2006), this time for a within-participant effect of time frame on confidence ratings. In their study, with syllogistic reasoning tasks, judgments provided under pressure to provide the first solution that came to mind were lower than later confidence ratings provided after participants could think freely, regardless of the extent of actual improvement in success rates between the two response stages. As these examples show, some task characteristics, like time frame, affect the perceived difficulty of the task, while others, like test format, going through training,

and the presentation medium, do not. Future research is called for to clarify which conditions people adequately take into account and which they ignore despite effects on performance.

Level 3: Momentary Experience

As mentioned above, the vast majority of meta-memory research dealing with heuristic cues for metacognitive judgments has focused on momentary subjective experiences that provide cues for item-level judgments. Yet, even with itemized judgments, there is room to consider which cues are theory-based, guided by people's beliefs regarding characteristics of the stimulus, and which cues are experience-based, and guided by gut feeling (Koriat, 1997; Mueller & Dunlosky, 2017; Undorf & Erdfelder, 2015).

A prominent experience-based heuristic cue in both meta-memory and meta-reasoning research is *processing fluency* – the subjective ease with which a cognitive task is performed. Processing fluency is typically measured by response time, and its utilization is indicated by a negative correlation between time and judgments. Overall, processing fluency is a valid cue for success (e.g., Koriat et al., 2006; see Unkelbach & Greifeneder, 2013, for a review). When solving an easy problem, people can come up with the right solution quickly and feel highly confident that their solution is correct. When facing a challenging problem, though, in many cases the chance for success remains low despite investing a lot of effort, and people acknowledge this in their confidence ratings, as found across domains (e.g., Ackerman & Zalmanov, 2012; Blissett, Sibbald, Kok, & van Merrienboer, 2018; Fernandez-Cruz, Arango-Muñoz, & Volz, 2016). Even feeling of rightness – a metacognitive judgment regarding initial intuitive answers that come to mind quickly – has been found to reliably reflect processing fluency (e.g., Thompson, Evans, & Campbell, 2013; Thompson, Prowse Turner et al., 2013).

The fact that negative time–judgment correlations are consistent across various research domains has been interpreted as indicating that fluency is a ubiquitous cue. However, this consistency does not rule out alternative explanations for the observed patterns. Ackerman (2014) suggested the Diminishing Criterion Model (DCM) as an alternative explanation for the negative time–judgment correlations. She called attention to the fact that fluency is a bottom-up inference process, whereby people first invest effort and then infer from the amount of effort already invested their chance for success at any given point (Koriat et al., 2006). According to the DCM, in contrast, people regulate their effort in a goal-driven manner, aiming to achieve a satisfactory chance for success (see Nelson & Narens, 1990). However, as they invest longer in each item, they compromise on their target level of confidence (that is, the level of confidence at which they will cease to invest effort). Thus, by the DCM, compromise generates the negative correlation between response time and judgments. This explanation does not rule out fluency as a heuristic cue altogether,

but suggests a combination of bottom-up inference and top-down regulation, which cannot be easily differentiated. In particular, Undorf and Ackerman (2017) found that the negative time-judgment correlation is limited to relatively high confidence levels (50-100%), while people invest a similar amount of time across all low levels of confidence (0-50%). According to the DCM, people stop when getting to a time limit, beyond which they are not willing to invest any further effort. Thus, it is possible that fluency have more effect when people feel knowledgeable and less so when they feel unconfident about their performance. This possibility deserves attention by future research.

Beyond response time, a number of other heuristic cues that predict confidence also have bearing on ease of processing. Memory research has found confidence to be positively correlated with three cues: consensuality of answers – the level of agreement across participants (Koriat, 2008); self-consistency – the consistency of the evidence supporting each answer option (Koriat, 2012); and accessibility – the number of associations that come to mind when answering a question. Confidence is also negatively correlated with cardinality – the number of considered answer options (Jackson, 2016). Meta-reasoning research supports and extends these findings. For instance, Bajšanski et al. (in press) found both consistency and cardinality to predict confidence in syllogistic reasoning tasks, even after controlling for response time. Similarly, Ackerman and Beller (2017), using solvable and unsolvable problems, found initial judgment of solvability to be associated with accessibility even after controlling for response time. Thus, although many cues are clearly associated with ease of processing, they often make a contribution beyond affecting processing speed.

A great deal of work has been done to identify conditions under which heuristic cues mislead judgments, and to expose factors that affect success rates but are not reflected in metacognitive judgments. To expose such biases, researchers triangulate confidence and accuracy with a measure that points to the heuristic cue under study. In the study mentioned above, Ackerman and Zalmanov (2012) had some participants solve problems using an open-ended test format, in which they had to type in their answer, while others solved the same problems using a multiple-choice test format, in which they had to choose the answer among four alternatives. Ackerman and Zalmanov triangulated confidence and accuracy with response time, as a measure of processing fluency (see another example of such triangulation in Shynkaruk & Thompson, 2006). A multiple-choice test format offers greater chances for success than an open-ended test format because solvers benefit from the opportunity to carefully consider each alternative, to recognize the correct answer when they see it, or even just to guess successfully; and as expected, solvers using that format had higher overall success rates. Ackerman and Zalmanov found that for all participants, confidence in each item was correlated with response time, presumably reflecting processing fluency. However, confidence ratings did not reflect the overall success rate difference between the test formats (see Figure 2).

Thus, as described earlier, test format was under-used as a cue for confidence judgments. This finding is in line with meta-memory findings that judgments of learning do not reflect appreciation of mnemonic methods that improve recall (e.g., imagining the memorized words, Rabinowitz, Ackerman, Craik, & Hinchley, 1982).

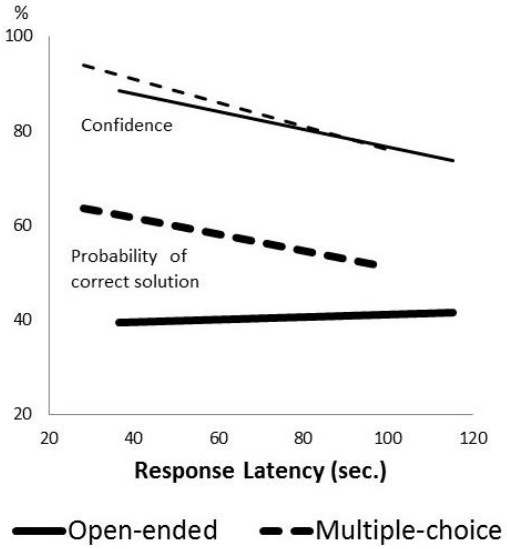

Figure 2. Confidence and probability of correct solution in an extended version of the Cognitive Reflection Test (see supplementary materials for Ackerman, 2014) presented in an open-ended or multiple-choice test format. Adapted from Ackerman and Zalmanov (2012).

Even more striking are findings indicating that people utilize heuristic cues which are in fact irrelevant for the task at hand. Topolinski (2014) reviewed a line of research in which participants were presented with word triads, of which half were solvable compound remote associates (CRAs) and half were random collections of three words. In a solvable CRA problem the answer is a common word that generates a compound word or phrase with each of the three (e.g., for the triplet FOOD – FORWARD – BREAK the correct answer is FAST, generating FAST FOOD, FAST FORWARD, and BREAKFAST). Participants had to decide quickly whether a word triad was coherent (solvable) or not. Topolinski and his colleagues (e.g., Topolinski & Strack, 2009) found that participants were more likely to judge words associated with positive affect as solvable, although the words' affective value was not associated with their solvability.

Misleading heuristic cues can lead metacognitive monitoring astray to the point where heuristic-based judgments and effects on success rates are in opposite directions. Topolinski, Bakhtiari, and Erle (2016) presented to participants solvable anagrams (scrambled words) and unsolvable letter sets which could not be rearranged to form a valid word, and manipulated their pronounceability. For

instance, the word EPISODE was turned into the pronounceable anagram EDISEPO and the less pronounceable IPSDEOE; similar alternatives were created for the unsolvable letter sets. As expected, easy-to-pronounce anagrams were rated as solvable more often than hard-to-pronounce anagrams, for both solvable and unsolvable anagrams. This finding is particularly interesting because in reality anagrams which are easier to pronounce are harder to solve, and indeed showed lower success rates, since people find it more difficult to rearrange the letters (Novick & Sherman, 2008). Thus, pronounceability is a misleading heuristic cue for metacognitive judgments.

Ackerman and Beller (2017) also used solvable CRA problems and random word triads. For each word in the triads, they used the number of compound words or two-word phrases in the language as an index for the heuristic cue of accessibility (Koriat, 1995). They then controlled for accessibility by balancing accessibility of the included words across solvable and unsolvable sets of problems, thereby making the frequency at which words create compounds or phrases an irrelevant cue for judging solvability. Notably also, the number of associations existing for a given word is not indicative of whether it shares a common association with the two words presented alongside it. CRA items containing words with a large number of associations on average across the three words are in fact harder to solve, rather than easier, because incorrect associations are more difficult to discount (in a manner similar to the effect of pronounceability on the difficulty of anagrams). Indeed, high-accessibility CRAs show lower success rates than low-accessibility CRAs. Nevertheless, Ackerman and Beller's participants judged problems containing words with high accessibility as more likely to be solvable than those where accessibility for all three words was low. Thus, unlike in memory contexts, accessibility of word triads is a misleading heuristic cue which is not only not indicative of solvability, but at odds with actual difficulty of the word triad.

Most of the heuristic cues considered in meta-memory and meta-reasoning research are based on semantic knowledge activated in verbal tasks, as is the case with pronounceability and accessibility of relevant knowledge reviewed above. Studying heuristic cues that affect performing non-verbal tasks provides opportunities to consider other types of heuristic cues. In a study by Boldt, De Gardelle, and Yeung (2017), participants judged the average color of an array of eight colored shapes and rated confidence in their choice. The greater the variability of colors across the eight shapes, the lower participants' confidence in their choice of the average color, even after controlling for the actual difficulty of the task. Reber, Brun, and Mitterndorfer (2008) found that symmetry was used as a heuristic cue when participants were asked to provide quick intuitive judgments about the correctness of dot-pattern addition equations. Lauterman and Ackerman (in press) manipulated original Raven Matrices to have unsolvable versions, by mixing the elements within each matrix, so to break the rules in the lines and columns. They presented participants a mixture of solvable and unsolvable matrices, balanced for

the original difficulty of each matrix. Participants had to judge quickly (4 seconds) whether the matrix is solvable in the first phase and to attempt solving in the second phase. This initial judgment of solvability, reported above to be predictive of later solving attempts, was associated with the original difficulty of the Raven Matrix, although, in fact, these two matrix characteristics, solvability and original difficulty, were unrelated. Thus, people utilize misleading heuristic cues in visual tasks as they do in verbal tasks.

In sum, metacognitive judgments are prone to predictable biases which stem from utilizing heuristic cues that are generally valid even in cases where these particular cues are misleading. Understanding what factors people take into account when making metacognitive judgments is important for knowing which conditions allow more attuned judgments and for guiding improvement attempts.

Methodologies for Exposing Heuristic Cues

Exposing a potential heuristic cue starts with proposing a factor that is expected to underlie a metacognitive judgment. The review above included some examples of heuristic cues (see Box 1). The findings, mentioned above, that people integrate multiple heuristic cues in complex ways (e.g., Bajšanski et al., in press; Undorf et al., 2018) hint that many cues are yet to be discovered.

A number of methodologies have been used to examine whether a suggested factor underlies a metacognitive judgment. Here I present the three main approaches emerging from the literature (see Box 2).

The main difference between methods aimed primarily at exposing valid cues (Method A) and methods useful for exposing biasing cues (Method B and Method C) is that the latter generate a dissociation between – or reveal that a given factor has a differential effect on – judgments and performance. This is of high importance, because when a judgment reflects performance differences reliably, it is impossible to identify with certainty a particular factor that generates this reliability. For this reason, Method A is the weakest of the three methodologies discussed here. In contrast, when judgments deviate from performance in a predictable manner, with an identifiable factor associated with the differential effect (Method B and Method C), we can draw stronger conclusions as to the contribution (or lack thereof) of this factor to the judgment. However, alternative factors that might also correlate with the bias must be considered and ruled out. I demonstrate each method by reviewing various examples, most of which have already been mentioned above.

Method A can be illustrated with findings that feeling of rightness, feeling of error, and confidence judgments are typically negatively correlated with response time (Fernandez-Cruz et al., 2016; Koriat et al., 2006; Thompson, Prowse Turner et al., 2013). Important for the current discussion is that in all these cases there are also negative correlations between actual success rates and response time. The negative time–judgment correlations are interpreted as pointing to fluency (operationalized as

response time) as a heuristic cue for the judgments. However, such findings mainly reflect differences in difficulty between items, which in turn can stem from numerous characteristics that influence judgments through various cues, some of them reviewed above (e.g., consensuality, accessibility, familiarity, etc.).

Box 2. Approaches to Identifying Heuristic Cues

A review of the metacognitive literature reveals three main approaches to exposing heuristic cues for metacognitive judgments. The methods are presented in order from the weakest to the most convincing, and with reference to whether the method is best suited to exposing a valid cue (Method A) or a biasing cue (Methods B and C).

Method A

Main objective: To expose a valid cue.

Approach: Identifying an association between different levels of the suggested factor and the judgment under investigation, in line with its effect on performance.

Method B

Main objective: To test utilization of a cue by exposing a bias in relation to particular task types or task items.

Approach: Showing that a factor differentially affects judgments and objective performance. The identifying characteristic of Method B is using different task items for each level of the examined factor.

Method C

Main objective: To test utilization of a cue by exposing a bias using identical task items.

Approach: Showing through manipulations of the examined factor that different levels of the factor differentially affect judgments and objective performance for the same items.

In the study by Ackerman and Zalmanov (2012) reviewed above, the relationships between fluency (operationalized as response time), confidence judgments, and success were tested for two test formats, multiple choice and open-ended. Thus, this study illustrates Method B. As described above, confidence in problem solutions dropped with time to a similar extent in both test formats (see Figure 2). However, there was an important difference in the validity of fluency in terms of actual success rates. Response time was a valid cue for the multiple-choice test format, where success rates dropped with time at the same rate as the confidence judgments, but not for the open-ended format, where no association was found between response time and chance for success: participants facing the open-ended format had a constant 40% success rate in all problems, both those where they responded quickly and those where they responded after lengthy thinking. Thus, this

study showed that confidence drops with time regardless of the actual association between solving time and chance for success. This is more convincing evidence suggesting that people utilize fluency as a cue for confidence than when judgment and success rates are affected similarly by the examined factor in all conditions. Ackerman and Zalmanov (2012) interpreted the literally identical pattern of time–confidence relationships in the two test formats as suggesting that people underestimate the effect of test format on their results and utilize fluency blindly, showing overgeneralization, even when fluency is not indicative of performance.

The weakness of Method B, demonstrated here by Ackerman and Zalmanov's (2012) study, stems from the pronounced effects of task difficulty on judgment accuracy. An alternative to the fluency effect as an explanation for the findings of Ackerman and Zalmanov (2012) with respect to the multiple-choice format is that the association between time and success rates reflects the greater ease of solving multiple-choice questions than open-ended questions, as indicated by the overall higher success rates for the former. By this reasoning, the fluency effect and task difficulty may have independently generated the observed pattern of confidence ratings for the open-ended and multiple-choice tasks, respectively, rather than blind utilization of the same heuristic cue.

Topolinski and Reber (2010), using Method C, provide even more convincing evidence for the role of fluency, operationalized by response time, in metacognitive judgments. They first presented to participants each problem, and then presented a potential answer – the target stimulus – after either a very short or a slightly longer delay (the short and long delays differed by 50 to 300 milliseconds). Participants had to judge whether the presented answer was the correct solution for the problem. For both correct and incorrect candidates, faster-appearing solutions were more frequently judged as being correct than those presented after a longer delay. The results were replicated with three different types of problem-solving tasks, showing the robustness of the phenomenon. Thus, this procedure rules out alternative explanations based on task difficulty as the source for the association between time and metacognitive judgment.

The studies just described are concerned with processing fluency, operationalized as response time. Thompson, Prowse Turner et al. (2013) examined another type of fluency: perceptual fluency, operationalized as font readability. Here, fonts were manipulated to be easier or harder to read, while the task and items remained the same; thus, this study also employs Method C. Thompson et al. found that font readability affected neither participants' judgments nor their performance (see Meyer et al., 2015, for a meta-analysis). The important contribution in this case is the distinction between types of fluency: response time, interpreted as processing fluency, was negatively correlated with both judgments and performance, while font readability, interpreted as perceptual fluency, was not correlated with either.

Two studies demonstrate how to transfer data gathered with less-convincing task designs to Method C, which yields more-convincing evidence, using a data

analysis approach. Markovits, Thompson, and Brisson (2015) compared deductive reasoning tasks phrased abstractly, using nonsensical terms (e.g., "If someone glebs, then they are brandup"), to the same problems couched in phrasing that was logically equivalent but concrete, using familiar objects and terms (e.g., "If someone cuts their finger, the finger will bleed"). Participants were then given a premise (e.g., "A person is brandup") and a set of conclusions (e.g., "The person glebs"), and asked whether the presented conclusions followed logically from the information given. Markovits et al. examined the effect of concreteness on judgments of solvability and final confidence. Because their manipulation changed the task stimuli, the study design was based on Method B. However, during their data analysis Markovits et al. controlled for differences in the accuracy of respondents' reasoning (i.e., whether their answers were logically correct), and found that above and beyond accuracy, final confidence was higher for the concrete versions of the problems than for the abstract versions. Thus, this study provides convincing evidence that concreteness (or familiarity) underlies judgments regardless of success in the task (see also Bajšanski et al., in press).

Ackerman and Zalmanov (2012) also controlled for accuracy in a second study reported in the paper described above. They used CRA problems, which generate a pattern of confidence that seems to be highly reliable – both confidence and accuracy drop as more time was invested in solving a problem (Figure 3, Panel a). To transform this Method A study into a more convincing Method C study, they divided the results data into correct and wrong solutions. This breakdown made time non-predictive of accuracy. After the breakdown, they still found negative time–confidence relationships independently for correct and wrong solutions (Figure 3, Panel b).

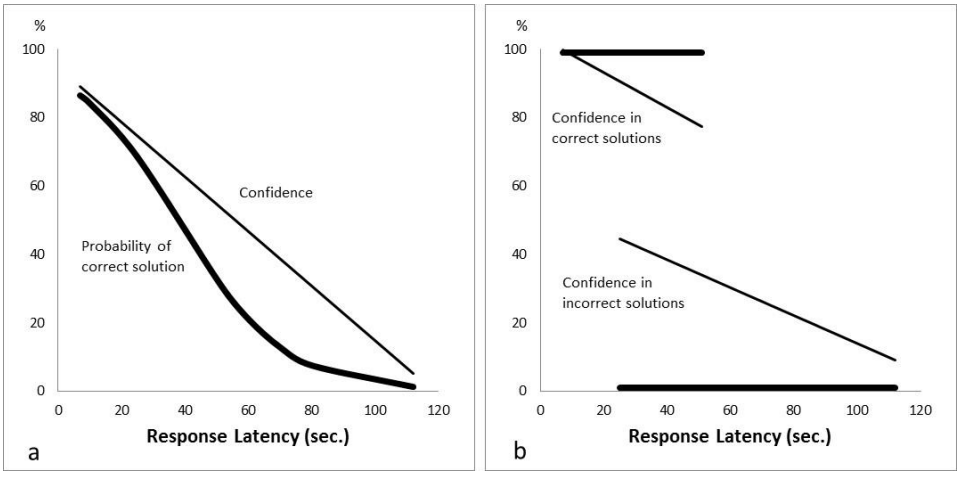

Figure 3. Confidence and probability of correct solution in a 30-item test with compound remote associates. Panel a: Overall pattern of results. Panel b: Breakdown of the same data into correct and incorrect solutions. Adapted from Ackerman and Zalmanov (2012).

Conclusion

Identifying the heuristic cues that underlie metacognitive judgments is the heart of metacognition as a scientific discipline, and in particular the meta-reasoning research domain. The methodological pitfalls reviewed in this paper make such research challenging. Through this paper I hope to help readers read the existing metacognitive literature more critically, and to support the design of high-quality research programs aimed at identifying and illuminating the heuristic cues that underlie meta-reasoning judgments. In particular, identifying heuristic cues allows us to expose conditions that may bias people's effort regulation – a necessary prelude to identifying conditions that support better performance in terms of both accuracy of results and efficient time management and guiding improvement attempts.

Cross-domain fertilization is also of high importance. A clear gap in the meta-reasoning literature is that although there are well-established methods for improving problem solving through educational support (e.g., see Frank, Simper, & Kaupp, 2018; Sweller, Merriënboer, & Paas, in press; Thibaut et al., 2018, for reviews), we do not yet understand how these methods affect metacognitive judgments in general and cue utilization in particular. For instance, problem solving can be improved when solvers accrue experience working with examples, and under some conditions this reduces overconfidence (Baars, van Gog, de Bruin, & Paas, 2014). From a metacognitive perspective, it is important to know whether this reduction in overconfidence is mediated by improved performance which is not reflected in confidence ratings; by a global decrease in confidence leading to an increase in invested effort; or by increased sensitivity to reliable cues that can be generalized to other contexts. Identifying the heuristic cues that people utilize spontaneously and those they can learn to use more effectively is a central goal of the meta-reasoning research domain.

Acknowledgements

I thank Monika Undorf, Tirza Lauterman, and Valerie Thompson for valuable feedback on early versions of this paper and to Meira Ben-Gad for editorial assistance.

References

Ackerman, R. (2014). The Diminishing Criterion Model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General, 143*(3), 1349-1368.

Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgments during reasoning and memorization. *Thinking & Reasoning, 23*(4), 376-408. doi:10.1080/13546783.2017.1328373

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18-32.

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior, 28*(5), 1816-1828.

Ackerman, R., & Thompson, V. A. (2015). Meta-reasoning: What can we learn from meta-memory? In A. Feeney & V. Thompson (Eds.), *Reasoning as memory* (pp. 164-182). Hove, UK: Psychology Press.

Ackerman, R., & Thompson, V. A. (2017a). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*(8), 607-617.

Ackerman, R., & Thompson, V. A. (2017b). Meta-reasoning: Shedding meta-cognitive light on reasoning research. In L. Ball & V. Thompson (Eds.), *International handbook of thinking & reasoning* (pp. 1-15). London: Psychology Press.

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency – confidence association in problem solving. *Psychonomic Bulletin & Review, 19*(6), 1187-1192.

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition, 146*, 377-386.

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*(3), 382-391. doi:10.1002/acp.3008

Bajšanski, I., Žauhar, V., & Valerjev, P. (in press). Confidence judgments in syllogistic reasoning: The role of consistency and response cardinality. *Thinking & Reasoning*, 1-34.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444.

Blissett, S., Sibbald, M., Kok, E., & van Merrienboer, J. (2018). Optimizing self-regulation of performance: Is mental effort a cue? *Advances in Health Sciences Education, 23*(5), 891-898.

Boldt, A., De Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance, 43*(8), 1520.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245-281.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131.

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review, 25*, 23-38.

Dentakos, S., Saoud, W., Ackerman, R., & Toplak, M. (in press). Does domain matter? Monitoring accuracy across domains. *Metacognition and Learning*.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science, 29*(5), 761-778.

Dunlosky, J., & Tauber, S. K. (2014). Understanding people's metacognitive judgments: An isomechanism framework and its implications for applied and theoretical research. In T. Perfect & D. S. Lindsay (Eds.), *Handbook of applied memory* (pp. 444-464). Thousand Oaks, CA: Sage.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98-121.

Fernandez-Cruz, A. L., Arango-Muñoz, S., & Volz, K. G. (2016). Oops, scratch that! Monitoring one's own errors during mental calculation. *Cognition, 146*, 110-120.

Frank, B., Simper, N., & Kaupp, J. (2018). Formative feedback and scaffolding for developing complex problem solving and modelling outcomes. *European Journal of Engineering Education, 43*(4), 552-568.

Jackson, S. A. (2016). Greater response cardinality indirectly reduces confidence. *Journal of Cognitive Psychology, 28*(4), 496-504.

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning, 9*(1), 25-49.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*(2), 157-175.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General, 124*(3), 311-333.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349-370.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 945-959.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*(1), 80-113.

Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (2014). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. *Journal of Experimental Psychology: General, 143*(1), 386-403.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*(1), 36-68.

Lauterman, T., & Ackerman, R. (in press). Initial judgment of solvability in non-verbal problems – A predictor of solving processes *Metacognition and Learning*.

Legg, A. M., & Locker, L. (2009). Math performance and its relationship to math anxiety and metacognition. *North American Journal of Psychology, 11*(3), 471-486.

Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & Cognition, 43*(4), 681-693.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174-179.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., … Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General, 144*(2), e16.

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 699-710.

Morsanyi, K., Busdraghi, C., & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: A potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions, 10*(1), 31.

Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language, 93*, 245-258.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125-173). San Diego, CA: Academic Press.

Novick, L. R., & Sherman, S. J. (2008). The effects of superficial and structural information on online problem solving for good versus poor anagram solvers. *The Quarterly Journal of Experimental Psychology, 61*(7), 1098-1120.

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning-Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review, 24*(6), 1774-1784.

Prowse Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking & Reasoning, 15*(1), 69-100.

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology, 37*(6), 688-695.

Reber, R., Brun, M., & Mitterndorfer, K. (2008). The use of heuristics in intuitive mathematical judgment. *Psychonomic Bulletin & Review, 15*(6), 1174-1178.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435-451.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34*(3), 619-632.

Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction, 51*, 61-73.

Stankov, L., Kleitman, S., & Jackson, S. A. (2014). Measures of the trait of confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 158-189). San Diego, CA, US: Academic Press.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747-758.

Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning, 19*(1), 54-77.

Sweller, J., Merriënboer, J. J. G. V., & Paas, F. (in press). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*.

Thibaut, L., Ceuppens, S., De Loof, H., De Meester, J., Goovaerts, L., Struyf, A., … De Cock, M. (2018). Integrated STEM education: A systematic review of instructional practices in secondary education. *European Journal of STEM Education, 3*(1), 02.

Thompson, V. A., Evans, J. S. B., & Campbell, J. I. (2013). Matching bias on the selection task: It's fast and feels good. *Thinking & Reasoning, 19*(3-4), 431-452.

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General, 147*(7), 945-961.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107-140.

Thompson, V. A., Prowse Turner, J. A., Pennycook, G., Ball, L., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*, 237-251.

Topolinski, S. (2014). Intuition: Introducing affect into cognition. In A. Feeney & V. Thompson (Eds.), *Reasoning as memory* (pp. 146-163). Hove, UK: Psychology Press.

Topolinski, S., Bakhtiari, G., & Erle, T. M. (2016). Can I cut the Gordian tnok? The impact of pronounceability, actual solvability, and length on intuitive problem assessments of anagrams. *Cognition, 146*, 439-452.

Topolinski, S., & Reber, R. (2010). Immediate truth-Temporal contiguity between a cognitive problem and its solution determines experienced veracity of the solution. *Cognition, 114*(1), 117-122.

Topolinski, S., & Strack, F. (2009). The analysis of intuition: Processing fluency and affect in judgements of semantic coherence. *Cognition and Emotion, 23*(8), 1465-1503.

Touron, D. R., Hertzog, C., & Speagle, J. Z. (2009). Subjective learning discounts test type: Evidence from an associative learning and transfer task. *Experimental Psychology, 57*(5), 327-337.

Undorf, M., & Ackerman, R. (2017). The puzzle of study time allocation for the most challenging items. *Psychonomic Bulletin & Review, 24*(6), 2003-2011. doi:10.3758/s13423-017-1261-4

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition, 43*(4), 647-658.

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition, 46*(4), 507-519.

Undorf, M., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(1), 97-109.

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In C. Unkelbach & R. Greifeneder (Eds.), *The experience of thinking: How the fluency of mental processes influences cognition and behaviour* (pp. 11-32). Hove, UK: Psychology Press.

van Loon, M., Destan, N., Spiess, M. A., de Bruin, A., & Roebers, C. M. (2017). Developmental progression in performance evaluations: Effects of children's cue-utilization and self-protection. *Learning and Instruction, 51*, 47-60.

Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied, 22*(4), 393-405.

Received: February 26, 2019